

DOCUMENT RESUME

ED 276 745

TM 860 695

AUTHOR Green, Donald Ross; Yen, Wendy M.
TITLE Content and Construct Validity of Norm-Referenced Tests.
PUB DATE Apr 84
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, New Orleans, LA, April 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Gains; *Achievement Tests; *Basic Skills; Construct Validity; Content Validity; Educational Objectives; Elementary Secondary Education; Mathematics Tests; National Norms; *Norm Referenced Tests; Skill Analysis; *Test Construction; Test Theory; *Test Validity
IDENTIFIERS *DMI Mathematics Systems

ABSTRACT

Basic skills tests in which measurement of growth is defined in the context of national norms are discussed. An integral part of the construction of norms is the production of a trait or ability score which is normed. Test publishers define a multidimensional trait by including items which measure it. The trait as a construct is a distinct ability, rather than the sum of a set of discrete skills. To be useful, a normative scale must exhibit monotonic growth. Curves which provide the basis for unambiguous norms are illustrated, using achievement data from a trial of the DMI Mathematics Systems, which is a set of criterion-referenced tests in mathematics suitable for grades K to 8 and above. Objectives were developed to be consistent with current classroom practices, curricula, and textbooks. Two test booklets, each containing 42 items, were tried out in grades 6-10, using over 2,000 students. The results for some skills showed decreasing characteristic curves from one grade to the next. It was concluded that items which result in non-monotonic increasing characteristic curves must be eliminated in order to have empirically based nonambiguous norms. When the norm-referenced trait involves growth, content areas that do not grow do not belong in its measurement. (GDC)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED276745

Content and Construct Validity
of Norm-Referenced Tests

Donald Ross Green

Wendy M. Yen

CTB/McGraw-Hill

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Donald Ross
Green

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

Presented at the annual meeting of the American Educational
Research Association, New Orleans, April 1984

TM 860 695

Content and Construct Validity of Norm-Referenced Tests

The issues being considered in this symposium have been around for a long time, say 50 years. These issues are simultaneously quite simple and quite complex depending on how they are stated, how broad or narrow the context in which they are discussed, and how many of them one is willing to consider simultaneously.

We prefer the simple approach partly because there is inadequate time to deal with the more complex interrelationships, partly because we are particularly interested in one simplifying and rather narrow context, and partly because our starting point is one relatively straightforward logical proposition and a set of illustrative data to go with it.

Our context is the measurement of academic achievement which we will limit further to the so-called "basic skills." Reading, mathematics, language are the common descriptors of these sets of skills. A second limitation is even more severe: we prefer to discuss measurement of growth defined in the context of national norms.

We will concede immediately that these limiting conditions remove the issues from the domains of interest of many people; our defense is that this is the area that interests us. It is also true that there is no absolute necessity to be concerned about growth; however we would note that growth is one very central concern to school personnel and thus worth substantial attention. We further concede that alternatives to the use of a comparative scale (i.e., norms) can be found.

We make the assumption that an integral part of the construction of norms is the production of a trait or ability score which is normed.

It is obvious that achievement in any of the basic skills is multidimensional. For example, what we call "computation" ability or skill is the combination or result of a large number of psychological processes. For the sake of practicality the tradition in measurement has been to produce a summary measure of an ability or trait that is then scaled and normed. We do not claim to have the definitive description of what any of these traits are, nor to our knowledge does anyone else. Every test publisher comes up with a somewhat different definition of the trait by the act of including different items to measure it. What the publishers have in common is that they all produce some summary measure of various basic skills. These basic skills are common constructs widely accepted in education; in fact measurement and curricula are built around these constructs.

We make the further assumption that although the measurement of these traits involves the simple addition of scores on discrete tasks (i.e., items), the trait as a construct is not the simple sum of a set of discrete skills. For example, "computation" ability is not the sum of the skills of adding single digit numbers, subtracting with regrouping, etc. The "computation" construct is the concept of a unified, developed ability to do computation. Knowledge of a specific computation fact is considered part of the trait only if it has been integrated into the broader skill. Part of our implicit assumptions with respect to the ability construct could be said to be overlearning; the ultimate result of learning computation skills is the development of a repertoire of skills that are well known, not easily forgotten, and integrated. Thus, the idea of growth and development is a fundamental aspect of our conception of the normed ability or trait construct.

In this narrow context then we would like to assert as a necessary, i.e., logical, proposition that to be useful a normative scale must exhibit monotonic growth. Figures 1, 2, 3 give examples of growth curves which do (Fig. 1) and do not (Figs. 2 and 3) meet this growth condition. Any one of the three curves in Figure 1 or other similar ones satisfy our criteria for the construct of a normed trait because they show growth.

The curves in Figure 1 provide the basis for unambiguous norms. Figures 2 and 3 on the other hand illustrate sequences of score changes that are unsatisfactory simply because there is no unambiguous interpretation of all scores on the growth scale. For example a score of Y in Figure 2c is characteristic of the average of grades 6, 8, and 10. Thus, a student scoring at Y has a score equal to that for the average of grades 6, 8, and 10, but is below the mean for grade 7 and above the mean for grade 9. In short the result is nonsense. Figure 3a/b is even more difficult to interpret at score Y. We may note that Fig. 3c is close to Figure 1c and that when flat growth occurs for extreme scores one can curtail the range of application of the tests producing scores on this scale to avoid such ambiguities.

It follows that if one wants a functioning test within our constraints it is necessary to delete or, rather, not include any sets of test items that exhibit one of these non-monotonic increasing growth patterns. Because it may not be obvious that non-monotonic increasing growth really occurs we offer some illustrative data.

These examples are taken from tryout data for the DMI Mathematics Systems (DMI/MS; Gessel, 1983). The DMI/MS is a set of criterion referenced tests in mathematics suitable for grades K to 8 and above. (Keep in mind that we are interested in norm referenced

tests; the fact that our examples of items come from a criterion referenced test is not important.) The major criterion for objectives to be eligible for the DMI/MS test is that their content be taught in the relevant grades.

The primary concern in defining the instructional objectives for DMI/MS was to ensure that the objectives would correspond as closely as possible with those taught in American classrooms. To this end, the first step in defining the objectives was to review the contents of 14 basal mathematics series, including some series already widely used and others, newly published, which were expected to become widely used....

This information was supplemented by a review of 18 curriculum guides....

The result of these studies was a list of some 250 separately defined objectives.... This list was next submitted to twelve reviewers [who] were asked, first, to indicate which were, in their opinions, the fifty most important objectives in the list; second, to indicate the grade(s) and semester(s) in which diagnostic testing of each objective should be performed; third, to indicate the grade(s) and semester(s) in which each objective is most commonly emphasized in instruction; and fourth, to indicate the grade(s) and semester(s) in which mastery testing of the objective should be performed.

The results of this survey, tabulated by objective and subobjective across grades and semesters, provided the

basis for selecting the DMI/MS tryout objectives and assigning them to levels (Gessel, 1984, p. 10).

The data collection for the tryout of the items was conducted in the fall of 1982. We are concerned here with two test books that were tried out at grades 6-10. Each test book contained three tests, with 14 items in each test. Test 1 in book 67 involved exponent items. Test 2 in book 67 was composed of addition and subtraction of integers items; five of these items involved integers with the same sign and nine items involved integers with opposite signs. Test 3 in book 67 included multiplication or division of integers; six of these were items with same sign integers and eight were items with opposite sign integers. Book 69 test 1 was composed of word problems involving fractions, and book 69 test 2 included word problems involving proportions. Book 69 test 3 had items related to prime numbers; nine of these items involved identifying prime numbers and five involved finding the prime factors for numbers.

The following numbers of students took Book 67 at the five different grades: 327, 515, 630, 411, and 472; book 69 was taken by the following numbers of students: 364, 331, 455, 598, and 405. These tryout groups were selected to be heterogeneous samples from across the country, but they did not constitute representative normative samples. Characteristic curves for the item sets were obtained. Each point in the characteristic curve is the average of the p values for five or more items at one grade. Given the sample sizes mentioned above, the standard errors of the points on the characteristic curves were all less than .02.

Because the tryout samples were not representative national samples, it was possible that irregular item characteristic curves could be obtained due to changes in the sample characteristics with grade. Therefore characteristic curves are presented for several tests within a tryout book. Decreasing characteristic curves that are found for one test or set of items but not others are evidence of decreasing performance that cannot be attributed to sampling peculiarities.

The characteristic curve for the exponent items in book 67 test 1 is in Figure 4a. These items show growth from grade 6 to 7 but not after grade 7. Figures 4b and 4c show patterns of growth that are similar: the opposite sign items show regular growth but the same sign items do not. In Figure 5a the word problems with fractions show no growth between grades 7 and 8, and the word problems with proportions show little growth from grades 6 to 9. In Figure 5b the prime factors items show little growth from grades 7 to 8, but good growth between all the other grades. On the other hand, the identification of prime numbers show decreasing performance from grade 7 to 8 to 9.

Consider what it would do to norms to include items with decreasing characteristic curves, such as those for same sign computations or identifying primes: in large numbers such items would lead to ambiguous normative data. Thus, if one wants empirically based norms, items with non-monotonic increasing characteristic curves must be eliminated at least from some of the grades regardless of one's convictions about content validity. The data lead one to argue that something is wrong with a normative achievement scale that would include items with decreasing characteristic curves.

One way to interpret items or objectives that do not grow is to say that they are not reinforced by other related learning until

later grades, i.e., such skills do not fit in with what else is being learned and therefore they do not grow. Thus, such non-growing skills are not part of the general mathematics ability being developed during these years, and they are not incorporated into the student's general body of knowledge and skill in this domain. Therefore the content does not belong in a norm-referenced test. It follows that to include such items in the test would water down the scale and obscure the growth that is occurring and reduce the construct validity of the test. Construct validity here is clearly superior to simple minded content validity unless one is indeed only interested in content mastery. If one believes that it is reasonable and useful to postulate a trait or ability in achievement, construct validity must prevail.

Messick (1975) discusses the importance of construct validity relative to content representation. In particular he refers to Loevinger's "substantive" validity, "the extent to which the content of the items included in (and excluded from?) the test can be accounted for in terms of the trait believed to be measured and the context of measurement" (Loevinger, 1957, p.661). Messick notes that substantive validity requires response or data analysis. "The exclusion of certain items from the test at this point on the basis of poor empirical properties may distort the test's representativeness in covering the domain as originally conceived. It is to be justified on the grounds that the resulting test is thereby a better measure of the trait..." (Messick, 1975, p. 961).

Yalow and Popham (1983) state, "If a test is constructed so that it constitutes a representative sample of the domain of interest, then we expect that an examinee's score on the test reflects how the examinee will perform on the domain of interest" (p.11). The norm

referenced trait is necessarily multidimensional and we argue that even though content appears to be an appropriate part of the domain, there can be "pockets" of skills or content that are not integrated with the main trait. In such a case the items measuring these content pockets, like those in Figures 4 and 5, do not give appropriate information about the trait, and the trait probably will not give appropriate information about the skill pockets. When the trait is a norm referenced one for which growth is important, content areas that do not grow do not belong in the measurement of the trait.

References

- Gessel, J. (1983). DMI Mathematics Systems. Monterey, CA: CTB/McGraw-Hill.
- Gessel, J. (1984). DMI Mathematics Systems Preliminary Technical Report. Monterey, CA: CTB/McGraw-Hill.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and education. American Psychologist, 30, 955-966.
- Yalow, E.S., & Popham, W.J. (1983). Content validity at the crossroads. Educational Researcher, October, 10-21.

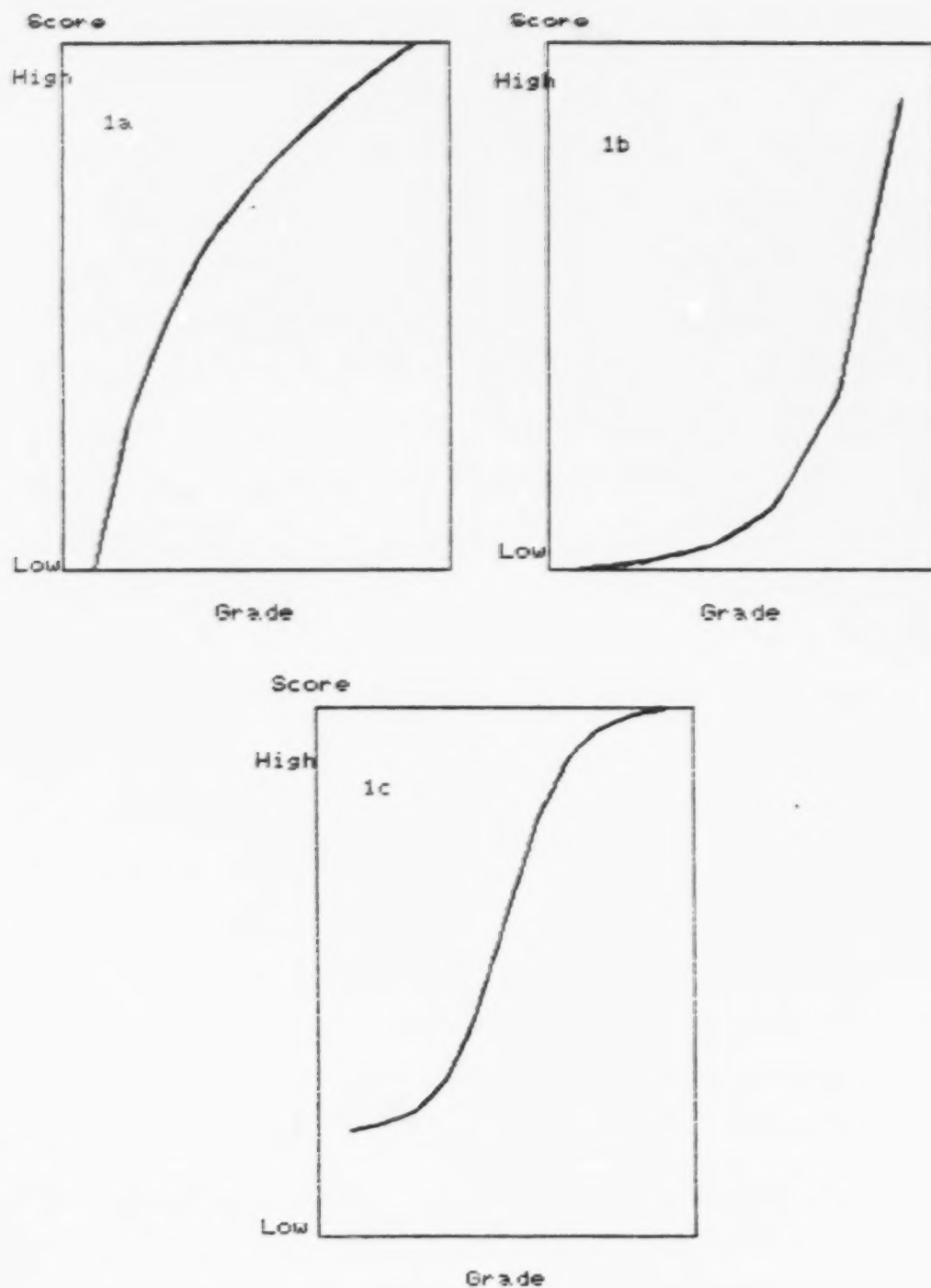


Figure 1. Examples of item characteristic curves showing monotonic growth.

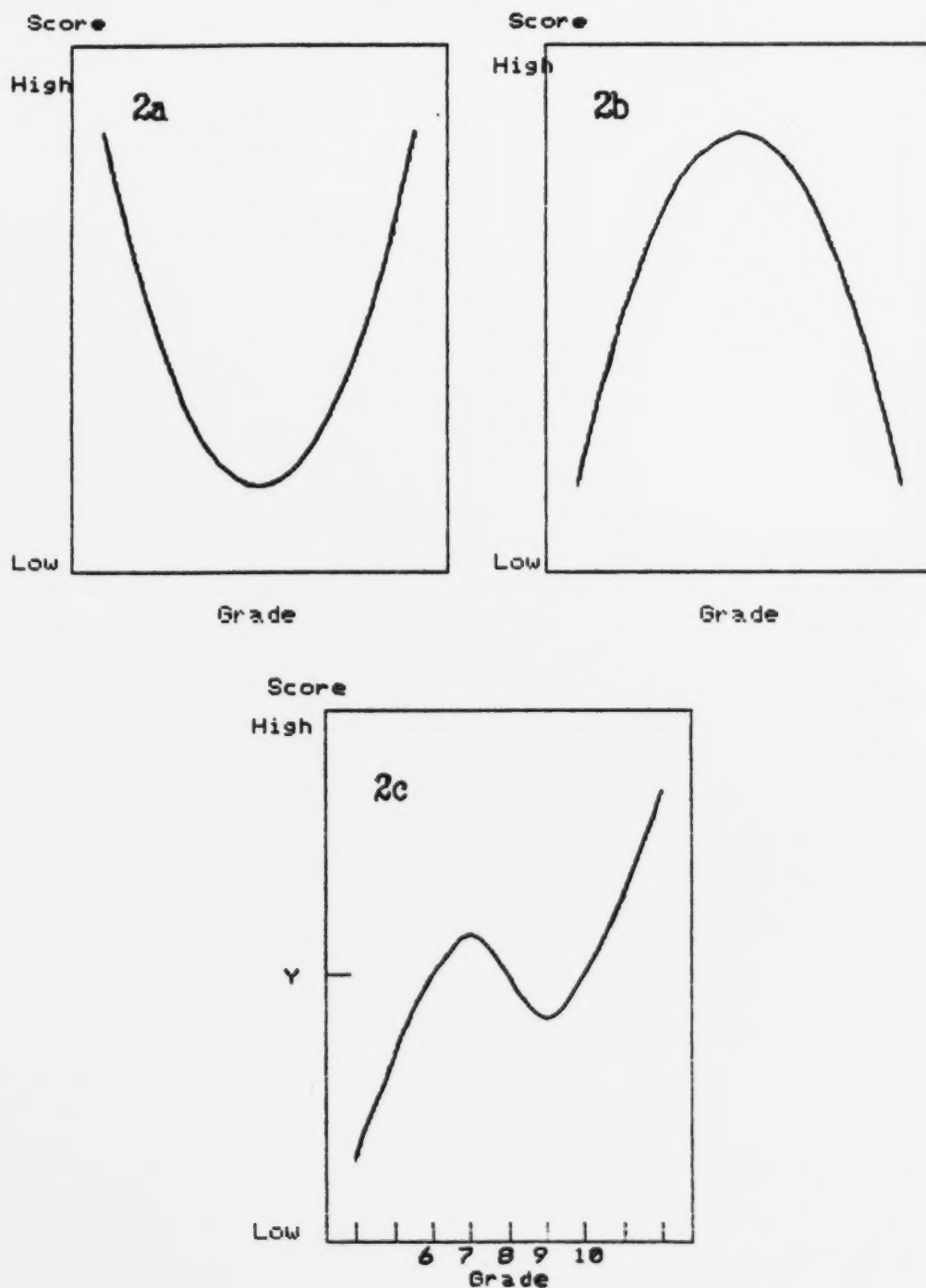


Figure 2. Examples of item characteristic curves not showing monotonic growth.

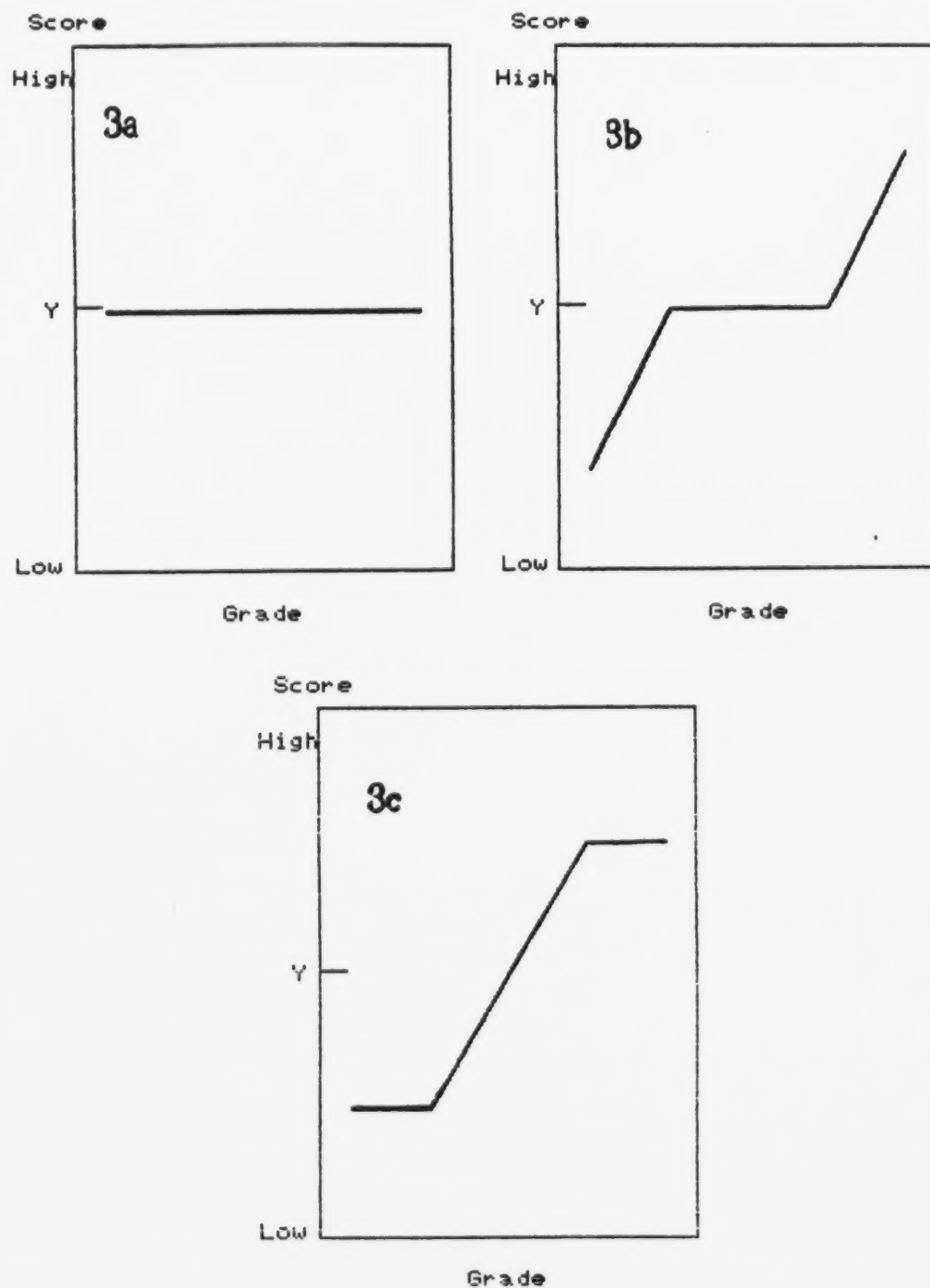


Figure 3. Examples of item characteristic curves that show areas of no growth.

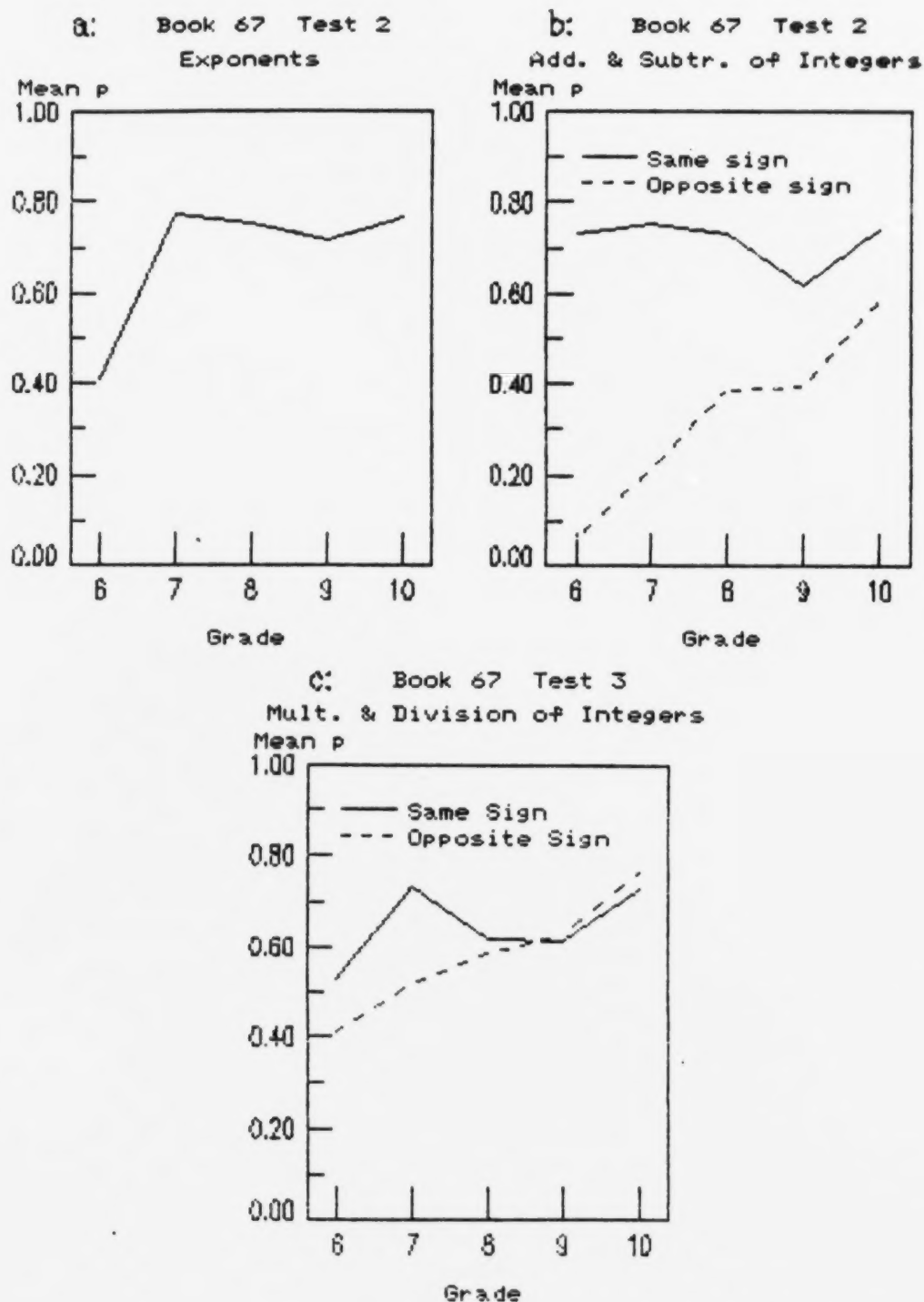
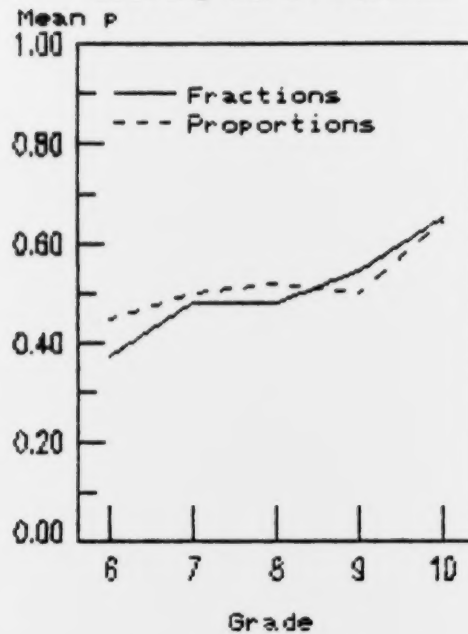


Figure 4. Mean p value by grade
for selected items: Book 67.

a: Book 69 Tests 1 & 2
Solving Word Problems



b: Book 69 Test 3
Prime Numbers

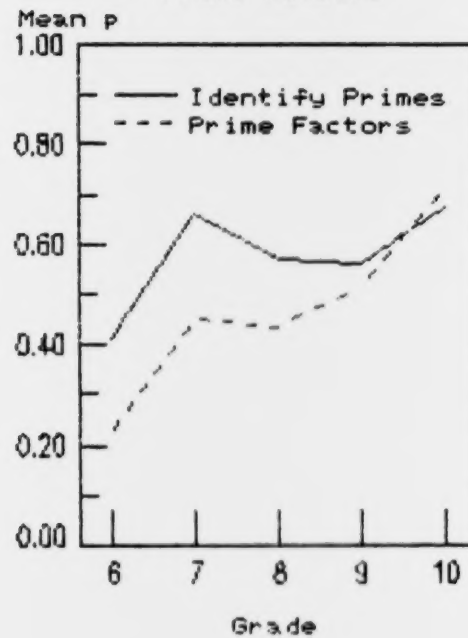


Figure 5. Mean p value by grade
for selected items: Book 69.